



Online Version Updated October 2009

U.S. DEPARTMENT
OF HEALTH AND
HUMAN SERVICES

National Institutes
of Health

Getting Connected with caBIG®

LIFE SCIENCES DISTRIBUTION

The caBIG® Life Sciences Distribution is a collection of software tools that facilitate life science research in support of a key goal of the National Cancer Institute (NCI)—to speed the delivery of innovative approaches for the prevention and treatment of cancer. The tools in this bundle enable management and analysis of diverse types of biomedical research data through a variety of capabilities, such as tracking and managing biospecimens, and analyzing microarray and imaging data.

All components of the caBIG® Life Sciences Distribution can be plugged into caGrid, the national infrastructure that virtually connects caBIG® informatics resources across and between research organizations. The individual tools may be downloaded at no cost and subsequently deployed or customized for use by organizations or networks.

This document provides an overview of the Life Sciences Distribution, its capabilities, its features and benefits, and the requirements for implementing the tools. More detailed information about individual tools can be found at <https://cabig.nci.nih.gov/adopt/LSD/>. A full list of caBIG® tools and capabilities is available online at <https://cabig.nci.nih.gov/inventory>.

Tools in this bundle include

- caTissue Suite—Biobanking management system
- Clinical Trials Object Data System (CTODS)—Virtual clinical data repository
- Cancer Genome-Wide Association Studies (caGWAS)—Genome-wide data management tool
- National Biomedical Imaging Archive (NBIA)—Image repository
- caArray—Microarray data management system
- geWorkbench—Tool enabling analysis, visualization, and management of gene expression, sequence, pathway, gene annotation and protein structure data
- caB2B—Search tool to identify data and services on caGrid
- caIntegrator2—Custom Web portals for integrative research
- caGrid—caBIG®-compatible systems architecture

Getting Connected with caBIG®

LIFE SCIENCES DISTRIBUTION



Tools	Description	Benefits	Capabilities	
caArray	caArray is a microarray data management system that guides the annotation and supports the exchange of array data and also connects with analytical tools like geWorkbench and caIntegrator 2.	<ul style="list-style-type: none">Provides Web browser-based and programmatic access to microarray dataSupports local management of microarray data with user control of collaborator and public access to samples and experimentsIntegrates data management and data analysis activities through interoperable tools and services	<ul style="list-style-type: none">Enables researcher submission and annotation of data consistent with MIAME guidelines using Web-based formsSupports bulk data import and export of MAGE-TAB filesProvides group-based permission scheme including settings for publishing data to the public domainEnables searches to readily discover and extract data of interestParses native data files from Affymetrix, GenePix, and Illumina platforms; stores and manages native files from many other microarray data formats	caArray
caTissue Suite	caTissue is a biobanking management tool that enables users to collect, manage, process, annotate, request, and distribute biospecimens and associated information.	<ul style="list-style-type: none">Provides Web browser-based and programmatic access to biospecimen dataManages tissue, fluid, cell, and molecular biospecimen informationAllows users to find and request specimens needed for molecular correlative studies	<ul style="list-style-type: none">Enables tracking of multiple specimens and refined materials (RNA, DNA, protein) and collection time points from a single participantUtilizes a role-based permission scheme for repository personnel and researchersFacilitates monitoring of quality assurance, distribution, derivation, and distribution data for primary biospecimens and derived materialsProvides a flexible storage container structure and biospecimen inventory management functions	caTissue Suite
geWorkbench	geWorkbench is a desktop bioinformatics platform that offers a comprehensive and extensible collection of tools for the management, analysis, visualization, and annotation of microarray-based gene expression and sequence data.	<ul style="list-style-type: none">Enables integrated access, analysis, and visualization of genomic data (gene expression, sequence, pathway, structure)Provides sophisticated analysis of genomic data through the integration of visualization tools, external databases, and computational services	<ul style="list-style-type: none">Provides analysis and visualization tools for microarray-based gene expression data from a variety of systems, including Affymetrix MAS5/GCOS, tab-delimited systems (e.g. GEO series matrix, RMAExpress), GenePix, and geWorkbench Matrix formatProvides analysis and visualization tools for gene and protein sequence data; plug-in components provide additional functionality for promoter, network, genomic DNA, protein structure, and pattern discovery dataSupports pathways (BioCarta), gene ontologies, networks, differential expression, clustering, and patterns based on regular expressions, as well as gene annotation retrieval from the NCI Cancer Gene Index and the Cancer Genome Anatomy Project (CGAP)Facilitates integrated access to external data sources and computational services (e.g., GoldenPath at Santa Cruz, NCBI BLAST)	geWorkbench
National Biomedical Imaging Archive (NBIA)	NBIA is a searchable repository of <i>in vivo</i> images, such as CT, MRI, and digital X-rays. NBIA also contains annotation files, including PDF and image markup provided by a curator.	<ul style="list-style-type: none">Serves as a platform for image data management and integration with other research data types, including clinical and genomic dataEnables development of imaging tools leading to improved clinical decision supportAccelerates diagnostic imaging decision making and quantitative imaging assessment	<ul style="list-style-type: none">Provides access to a searchable repository of biomedical images along with key annotations, as well as image archives and imaging resourcesFacilitates development and validation of analytical software tools that support lesion detection and classification, accelerate diagnostic imaging decisions, and quantify imaging assessment of drug response	NBIA
caBench-to-Bedside (caB2B)	caB2B allows researchers to easily and securely search for data across virtually any caGrid data service and join diverse datasets based on user-specified selection criteria. The caB2B 3.1 Suite of tools contains the caB2B Web Application, the caB2B Client Application, and the caB2B Administrative Module.	<ul style="list-style-type: none">Allows users to search for data across caGrid, enabling integrative research and data analysis	<ul style="list-style-type: none">caB2B Web Application: Enables keyword and form-based queries to retrieve data across the grid without needing to understand how the data is stored, through an easy-to-use Web search toolcaB2B Client Application: Enables advanced end users to create, save, and execute queries across caGrid data services and allows users to save retrieved data as a “virtual experiment” and visualize data using various graphical componentscaB2B Administrative Module: Enables users to customize caB2B instances by selecting domain models and service instances for queries, creating new categories, curating paths between classes, and creating inter-model joins	caB2B
Cancer Genome-Wide Association Studies (caGWAS)	caGWAS allows researchers to integrate, query, report, and analyze significant associations between genetic variations and disease, drug response or other clinical outcomes.	<ul style="list-style-type: none">Facilitates rapid sharing of information and results analysis from various biomedical studiesAllows researchers and bioinformaticians to access and analyze clinical and experimental data across multiple clinical studies	<ul style="list-style-type: none">Utilizes a standardized model to represent SNP genotype data, SNP association findings, population frequency data, and clinical phenotype dataProvides support for search and retrieval of GWAS findings in the context of genes or chromosomal regions of interest <p>NOTE - Due to recent developments in forensic science, it is recommended that human specimen data stored in caGWAS not be connected to caGrid.</p>	caGWAS
Clinical Trials Object Data System (CTODS)	CTODS is a database and software system for storing and sharing clinical trials data in both identifiable and de-identified form.	<ul style="list-style-type: none">Enables research organizations to utilize data from any in-house, caBIG®-compatible Clinical Trials Data Management System (CDMS) or data source for non-clinical researchSimplifies sharing of or access to de-identified clinical trials data over caGrid	<ul style="list-style-type: none">Uses open standards and standards-based tools designed to enable the cancer research community to share, interpret, and integrate de-identified informationLeverages the Biomedical Research Integrated Domain Group (BRIDG) model that underpins data interchange standards and technology solutions that enable harmonization between the biomedical/clinical research and healthcare arenas	CTODS
caIntegrator2	caIntegrator2 is a framework that allows researchers to create custom caBIG®-compatible Web portals to bring together heterogeneous clinical, microarray and medical imaging data, thereby facilitating integrative research.	<ul style="list-style-type: none">Allows researchers to execute, save, and share complex queries across multidimensional data setsEnables remote access to multiple caGrid-compatible data sets by integration with caArray, NBIA, caBIO (cancer Bioinformatics Infrastructure Objects), and GenePattern (platform for integrative genomics)	<ul style="list-style-type: none">Enables users to conduct Web-based study deployment management without programming experienceFacilitates mapping of study annotations to caDSR (Cancer Standards Data Registry and Repository) data typesLeverages existing grid capabilities for data analysis through integration with GenePattern and BioConductor (genomic data analysis software)	caIntegrator2
caGrid	caGrid is the service-oriented, federated architecture that connects caBIG®-compatible systems together regardless of location to facilitate data sharing and collaboration.	<ul style="list-style-type: none">Facilitates queries across disparate data resources located in geographically distributed locationsEnables integration of comparable data from different sourcesUses grid-wide resources to create workflow pipelines for data retrieval and analysis	<ul style="list-style-type: none">Connects caBIG®-compatible systems via a broadly deployed, standards-compliant data and analysis gridUtilizes a Business Process Execution Language (BPEL) workflow engine to enable data analysis pipeline construction and executionEnables federated, cross-domain data mining and integrationUtilizes standardized programming interfaces to simplify creation of interoperable applications	caGrid